Diversifying Search Results through Pattern-Based Subtopic Modeling

Wei Zheng, Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA

Hui Fang, Department of Electrical and Computer Engineering, University of Delaware, Newark, DE, USA

Hong Cheng, Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, Shatin, NT, Hong Kong

Xuanhui Wang, Facebook, Menlo Park, CA, USA

ABSTRACT

Traditional information retrieval models do not necessarily provide users with optimal search experience because the top ranked documents may contain excessively redundant information. Therefore, satisfying search results should be not only relevant to the query but also diversified to cover different subtopics of the query. In this paper, the authors propose a novel pattern-based framework to diversify search results, where each pattern is a set of semantically related terms covering the same subtopic. They first apply a maximal frequent pattern mining algorithm to extract the patterns from retrieval results of the query. The authors then propose to model a subtopic with either a single pattern or a group of similar patterns. A profile-based clustering method is adapted to group similar patterns based on their context information. The search results are then diversified using the extracted subtopics. Experimental results show that the proposed pattern-based methods are effective to diversify the search results.

Keywords: Clustering, Diversity, Frequent Pattern Mining, Information Retrieval, Subtopics

INTRODUCTION

It has been a long standing challenge to develop effective retrieval models that can provide users with optimal search experiences. Traditional retrieval models rank documents based on only their relevance scores and ignore the semantic relations among returned documents. However, different documents may contain the same piece of relevant information, and returning all of them would not be a good ranking strategy. Intuitively, search results covering different pieces of relevant information, i.e., query subtopics, are more desirable than those covering the same
single piece of relevant information multiple times. Thus, it is necessary to rank documents based on not only relevance but also diversity.

Diversifying search results can benefit both queries with extrinsic diversity such as ambiguous queries and those with intrinsic diversity such as under-specified queries (Clarke, Craswell, & Sobooff; Craswell, Fetterly, Najork, Robertson, & Yilmaz; Radlinski, Bennett, Carterette, & Joachims). The general goal of result diversification is to return a list of relevant documents that cover all of the subtopics of a query. As an example, query “java” is ambiguous. Since this query could have several interpretations and we may not know which interpretation reflects a user’s need, it would be important to return diversified results covering different interpretations so that they can satisfy all possible information needs. Another example is that a user doing a literature survey uses query “computer architecture” to find documents that cover representative topics in computer architecture. The user would prefer a ranking of documents covering different topics in computer architecture while avoiding excessive redundancy. The state of the art diversification methods aim to diversify search results so that they can cover more query subtopics. Thus, one of the key challenges is how to identify semantically meaningful subtopics for a given query.

In this paper, we propose to use the frequent pattern mining approach to identify query subtopics from a document set. We design a novel result diversification framework that models the diversity explicitly through pattern-based subtopics. The basic idea is to combine the relevance, through existing relevance-based retrieval models, with the diversity, through pattern-based subtopic modeling. In particular, we define a pattern as a set of semantically related and meaningful terms extracted from documents. For example, a pattern could be a phrase or a term collocation such as “programming language”, “code developer” and “gourmet coffee”. Such patterns can be efficiently discovered with the state-of-the-art frequent pattern mining algorithms (Bayardo, 1998; Han, Pei, & Yin, 2000; Zaki, 2000). We propose to model a query subtopic by a single pattern that is relevant to the query. However, since different patterns could be semantically related and complementary, they can be merged to form a more complete semantic unit. Thus, as an alternative, we model a query subtopic as a group of semantically related patterns that are relevant to the query. For example, by grouping the two patterns “programming language” and “code developer”, the subtopics of query “java” could be “programming language code developer” and “gourmet coffee”. To discover the related pattern groups as query subtopics, we use a profile-based pattern clustering method to group semantically related patterns (Yan, Cheng, Han, & Xin, 2005). The method was originally designed to summarize frequent itemsets into different groups so that similar item sets are assigned to the same group. The similarity between two patterns is measured based on not only the pattern composition, i.e., terms contained in the patterns, but also the context of patterns, i.e., documents containing the patterns. In this work, we represent the context of a pattern with a profile, which is the term distribution of the document set that contains the pattern. The similarity between two patterns can then be measured by the divergence between their profiles and similar patterns are grouped together to model one query subtopic.

Our main contribution is the novel methods to model query subtopic based on a frequent pattern mining approach, which effectively identifies query subtopics based on term co-occurrences. The pattern-based subtopic modeling methods allow us to focus on the important content of the documents and are more robust to the noises in the documents. To the best of our knowledge, no existing work on search result diversification used the pattern-based subtopic modeling idea. Furthermore, it provides a demonstration of how data mining techniques, in particular, frequent pattern mining, can help solve information retrieval problems.

The rest of the paper is organized as follows. We first discuss related work and present an overview of the proposed pattern-based
A New Web Site Quality Assessment Model for the Web 2.0 Era
Minseok Pang, Woojong Suh, Jinwon Hong, Jongho Kim and Heeseok Lee (2010). 
*Handbook of Research on Web 2.0, 3.0, and X.0: Technologies, Business, and Social Applications* (pp. 387-410).
[www.igi-global.com/chapter/new-web-site-quality-assessment/39182?camid=4v1a](www.igi-global.com/chapter/new-web-site-quality-assessment/39182?camid=4v1a)

A New Similarity Measure for Automatic Construction of the Unknown Word Lexical Dictionary
[www.igi-global.com/article/new-similarity-measure-automatic-construction/4110?camid=4v1a](www.igi-global.com/article/new-similarity-measure-automatic-construction/4110?camid=4v1a)